UCRL-JC-151404

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Validation, Uncertainty, and Quantitative Reliability at Confidence (QRC)

R. Logan, C. K. Nitta

December 06, 2002

American Institute of Aeronautics and Astronautics,
Reno, NV, January 5-9, 2003

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

VALIDATION, UNCERTAINTY, AND QUANTITATIVE RELIABILITY AT CONFIDENCE (QRC)

Roger W. Logan
Member, AIAA
University of California
Lawrence Livermore National Laboratory, L-125
Livermore, CA 94551

Cynthia K. Nitta
University of California
Lawrence Livermore National Laboratory, L-096
Livermore, CA 94551

ABSTRACT:

This paper represents a summary of our methodology for Verification and Validation and Uncertainty Quantification. A graded scale methodology is presented and related to other concepts in the literature. We describe the critical nature of quantified Verification and Validation with Uncertainty Quantification at specified Confidence levels in evaluating system certification status. Only after Verification and Validation has contributed to Uncertainty Quantification at specified confidence can rational tradeoffs of various scenarios be made. Verification and Validation methods for various scenarios and issues are applied in assessments of Quantified Reliability at Confidence and we summarize briefly how this can lead to a Value Engineering methodology for investment strategy. Due to the evolving nature of such methodologies, this work represents the views of the authors and not necessarily the views of Lawrence Livermore National Laboratory.

INTRODUCTION:

A process that leads through Verification & Validation (V&V) and eventually to the investment strategy is shown in Figure 1. First we establish system environment and model Requirements. Based on these, the V&V process leads to models with Uncertainty Quantification (UQ) at confidence [C]. We use these models to obtain margins [M] and reliability [R] equivalents, and Quantified Reliability at Confidence (QRC). We use QRC products for Value Engineering.

A Value Engineering related Quantified Systems Value (QSV_0) is defined and then adjusted as a function of Reliability at Confidence (RC*) over the system environments (i=1,E), with the economic function of Present Value Factor (PV_F) in the time (t) domain:

$$\square QSV = QSV_0 \int PV_F[t]\,\Delta[t]\{\Pi_{i=1,E}\ (RC^*)_i\}\,dt \qquad ......[1]$$

Key to the investment strategy process, and its linkage back to V&V, is the Benefit/Cost Ratio (BCR): Benefit (B) is proportional to Risk Reduction, expressed as ΔQSV. With the above relation, ΔRisk is proportional to Δ(QSV), and hence to Δ(QRC), and finally Δ(QRC) links to the fidelity of our Quantitative V&V statements (confidence bounded uncertainties). Then "BCR" is (Benefit-Cost)/Cost. Quantified V&V will show us that there is not a unique BCR – we must explore its bounds for any given decision. Our decisions will fall into 3 basic bins:

1. High BCR within our V&V bounds: Positive decision indicator [i.e. "do it"]
2. Low BCR within our V&V bounds: Negative decision indicator [i.e. "don't do it"]
3. BCR varies high to low depending on V&V bounds: [i.e. more quantification is needed]

The end product and dollar benefit can be explained using a Risk=Likelihood*Consequence Matrix as shown in Figure 2. V&V plays a quantified role, one that is now directly proportional to Risk Reduction and Value Engineering quantities.

Requirements and Graded Scale V&V

The first step is to establish system (and hence model) requirements. Based on these requirements for the system and its environments, the V&V process begins, leading to validated models with uncertainties at confidence. To enable a semi-quantitative V&V evaluation on our way to a quantitative V&V statement,

we are developing a continuously evolving 19-point VERification and 35-point VALidation checklist, with suggested criteria to consider when performing V&V analyses. These factors are summarized in an overall 10-point summary.

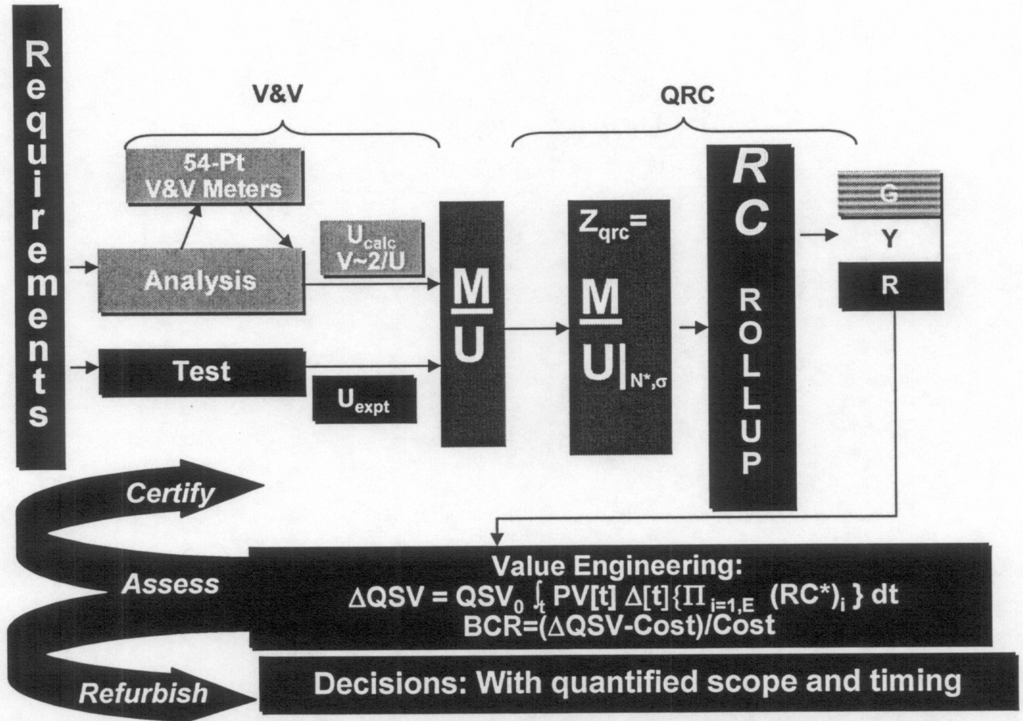


Figure 1. Flow diagram from system Requirements through V&V, through uncertainty quantification and margins; onward through QRC, then Value Engineering, QSV, and Benefit/Cost Ratio BCR.

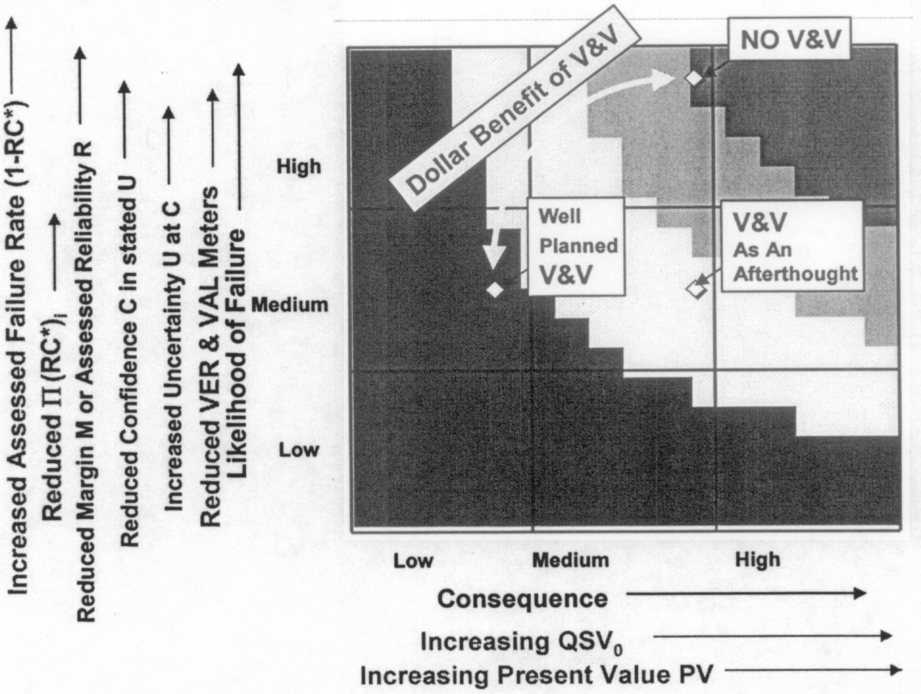


Figure 2. Dollar Benefit of V&V and Quantitative Certification, expressed as a standard Risk=Likelihood*Consequence Matrix. Likelihood becomes analogous to assessed (1-QRC); Consequence is expressed in Value Engineering / Earned Value [ie dollars] terms.

DEFINITIONS of VERIFICATION and VALIDATION:

We begin by describing our favorite V&V definitions and a V&V process that leads to qualitative measures for V&V. Following the *qualitative* process enables *quantitative* Validation Statements expressed with Uncertainties (U) at Confidence [C]. A compendium of definitions proposed from the community is available in the literature. Our preferences include the following:

Verification:

Verification[1]: "Verification of a CODE: The process that determines that the computer code accurately represents the mathematical equations".

Verification[1]: "Verification of a CALCULATION: The process that determines that the computer calculation for a particular problem of interest accurately represents the solutions of the mathematical model equations".

These definitions should enable a *quantitative* Verification Statement, such as the following example:

"This material model feature has demonstrated 99.3% accuracy on elastic plastic deviatoric stresses and strains [supporting documents should exist and be cited], and it has shown this accuracy in combination with 8 element types with aspect ratios as high as 5 and angles as low as 50 degrees. The model is known not to work well with values of bulk-to-shear modulus higher than 10. The model has shown over 96% accuracy with contact bulk stiffness ratios as high as 1000."

Validation:

Validation[2]: "The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model." (*Quoted from Oberkampf and Trucano[3]).*

Validation[1]: "Validated Model: A model that has confidence bounds on the output. A validated model output has the following characteristics:

1. The quantity of interest
2. An estimate of the bias
3. A set of confidence bounds.

A validated model is one where we can make a formal statement after running the model similar to:

'I am 90% confident that if I build and measure the quantity of interest, that it will fall within the confidence bands (of uncertainty) shown around the model output.'"

General Statement of Methodology:

We suggest evaluation of the code and model simulation results at three layers of comparison analogous to that suggested by AIAA[2], as well as a Software Quality Engineering (SQE) Layer "0" that supports the confidence in the software engineering underlying all of the other V&V activities.

All of these layers, including Layer 0 SQE, must be done using a graded scale proportional to the risk, consequence, and investment in the codes and models anticipated.

LAYER 0:

The V&V Program must first address evaluation of the SQE Practices and improvement for each code effort. SQE is a broad topic not addressed in this short paper.

LAYER 1:

The first layer compares simulation results to known analytic and semi-analytic solutions for specific verification test problems (Verification).

LAYER 2:

The second layer compares simulation results code versus code in a regime of interest that is known to be verified and validated to a given degree. Code vs. code "Validation by Similarity" is certainly not rigorous, but it may be cost effective, allowing us to save precious resources for more rigorous V&V of higher risk environments. Many note the cautions of optimistic results from code-to-code comparison. Code-vs-code validation can *never prove that the pair of code/model combination is right* – but it can prove that *at least one of them is wrong* – and by how much. Often this can be done with orders of magnitude less effort than physical validation testing. Code-to-code comparisons offer partial remedies to the following:

- Some tests cannot be done due to treaty, cost, lack of parts, funds, health and safety
- Comparing to legacy data is indeed essential, but often not as convincing as we would like: Our number of data points "N*" is small for a given system, and the detailed data is usually lacking – and indeed if it had been obtained we would encounter at length the dilemma that "you cannot take a measurement without perturbing the result of your measurement"
- Code-to-code validation, while risking

interdependency, allows comparison of nearly any desired quantity in any desired level of detail – at nil additional cost or fidelity risk

- Legacy vs. modern codes can often be compared – we can make statements about "validation compared to the legacy validation level" – if we consciously include any new "degrees of freedom (K)"
- In an era where computation cost is falling and test cost is rising, a strong V&V code-to-code element is compelling from an efficiency standpoint – before proposing expensive tests with scarce resources and competing priorities and liability risks.

LAYER 3:

The third layer compares model simulation results against legacy and modern experimental data (Validation). The legacy comparisons can be properly termed either **Calibration** (involving tuning of the model), or **Postdictive Validation** (using the model with no more tuning, as if the data were new). The modern test comparisons can be properly termed **Predictive Validation,** if we take advantage of these rare and expensive opportunities to make a prediction before our new tests are run. Of course, the Layer 3 process is iterative: We can repeat our Calibration process, or expand it at the expense of the old data remaining for Postdictive Validation. However, in doing so, we raise the **CAL/VAL Ratio, or the ratio of test data we have used to "tune" our models to the test data we use without tuning**). We can then sharpen our predictions and obtain, with yet more new tests, an improved assessment of Predictive Validation ability.

Of necessity intertwined with the three layers for code comparison within the V&V Program, Uncertainty Quantification (UQ) is essential as an integral focus area for evaluating the many sources of error in simulation, including uncertainties from the experimental databases, the code algorithms, software implementation, and physical and material models. Due to our preferred definitions of Validation above, we view V&V as requiring "ensemble computing" as stressed by Oberkampf and Trucano[3]. That is, the UQ process involves extensive amounts of model runs for sensitivity and uncertainty studies to enable quantitative validation statements that must accompany any claims of a level of validation.

Methodology and Independence Metrics

Ideally V&V would be a "one-pass" process, where codes and code features would be passed to the independent design teams, V&V'd, and cleared for production use. We postulate this for simplicity – but of course in reality V&V is an iterative process with the code teams, as both the developers and the users verify both the codes and the familiarity with them. We enable that iterative process by running combinations of the code-team supplied "regression" or "example" or "verification" suite – *but for formal V&V even some of these verification runs are made by the V&V team – the people outside the code development team – who are also the eventual team of designer / analysts.* In this way, we assure a product ready for Validation, where the "product" is a system of the code, user, documentation, support, and platform. This is not a slight to the code developers whatsoever – many of those in V&V have been developers at one time and recall the dilemma that when one develops a piece of code, it is common that the developer can successfully use the capability – but other users cannot. *Verification and Validation cannot be formally declared until a user base (the designer/analyst groups) declare it so.*

Verification and Validation Meters and Checklists

We can envision at least 4 methods for expressing the pedigree of V&V for a given code feature or model. Consider the following 4 methods:

A. Use of words like "Fully Validated", "Unvalidated", "Validated Code", etc. *(We do not recommend this method.)*

B. Use of a simple one-number "scale" e.g. 0-1, 0-5, or 0-10 as we suggest in our description of our "0-10 Scale" VER and VAL Meters. The VER and VAL Meters provide some measure of acceptability and caution; they are quantitative but they are subjectively set.

C. A multi-point "checklist" of thought, procedure, and documentation – including partial completion of (D) as well.

D. Quantitative error and uncertainty bounds at confidence, over a regime and quantity of interest for design and decision purposes.

We wish to determine what, of (A-D), should be a minimum set, and to express the level of V&V most efficiently and yet adequately, and as quantitatively as possible yet in a compact form. Full use of method (C) or (D) may not be warranted. However, use of the descriptors in (A) conveys very little meaning. We suggest that as a minimum, method (B) – a simple "scale" – conveys an expert opinion rating of the V&V level of a simulation in a condensed form. Certainly this is not a complete method, yet it conveys a more accurate picture than (A), and is thought provoking enough to lead us to pursue methods (C) and (D).

As we will show and see in the discussion below, we can quantify metrics but to have much meaning, and to specify an acceptance level, it is usually necessary to also answer "how much difference it will make if we are inside, at, near, just outside, or way outside" a given level. ***We assert that we can and should use methods consistent with a statistical quantification of reliability at confidence,*** *even though we cannot commonly use the frequentist approach but must rely on expert system, fuzzy logic, or Bayesian sets as described by Oberkampf and Trucano*[3] *and Booker et al*[4]. Methods described by Logan and Nitta[5] can eventually help us with "acceptance" levels for Validation. Uncertainties and Sensitivities in the simulations can be quantitatively related to *assessed levels* of Confidence and Reliability numbers for the stockpile during this era of no nuclear testing.

The goal of ***Validation*** is to take the model to a quantified state where they can be evaluated for acceptance for assessment and certification work at a demonstrated level of uncertainty and hence confidence. Our analysis timelines span many orders of magnitude – sometimes a "hero" fidelity simulation is best, but may require 2 years to build and run. Other times, an answer just incrementally better than back-of-the-envelope is needed, but that answer is needed in 2 hours. Balance of *sufficiency and efficiency*[6] is the key – balance of funding, timelines, priority, and credibility. Tradeoffs are always necessary between these. We simply must know – and express – what tradeoffs we have accepted.

The Qualitative VER and VAL Meters

One basic dilemma is to express the "V&V" pedigree of a simulation result or conclusion, in a way that goes beyond "yes or no", but remains a fairly simple quantitative expression of the V&V status of a simulation. Simplicity is essential to the decision making process, because calculation results and movies are often shown at fast paced meetings where numerous topics are covered in a few hours – with only cursory detail and never enough time for the audience to evaluate the credibility of the detail being shown. Typically only a few seconds are available to describe the pedigree of a given part of the simulation. And yet, impressions are formed at such meetings and can lead to misunderstandings and regrettable decisions, unless *at a minimum* some kind of graded scale V&V measure is used. There is nothing wrong with using a lower quality or conceptual analysis to make a point or point out an area of risk. However, to avoid having the audience take such examples with verbatim precision, a VER and VAL meter or equivalent *as a minimum* should be used.

But, the meters are of course relevant for more than just a "quick indicator" at fast paced review meetings. The Meter readings (or any such rolled-up number rating for V&V) can, in addition:

- Enhance the capability of "designer-centric" or "expert judgment based" V&V;
- Firm up the credibility of conclusions that are drawn using any historic methodology;
- Make more scientific the V&V process;
- Make more scientific the decision process;
- Provide fundamentals for rational discourse on this subject;
- Provide a rational basis for common understanding and expression of V&V level.
- Provide an expression of *relative* information and level regarding V&V

The concept of a Verification "Meter", with a Scale that reads 0-10, is simple but it should be fairly clear in intent. We suggest that the following factors help in setting this somewhat subjective but informative "0-10" Verification Scale:

Reading of 0-1.5:
- Code has a name and user's manual
- Version Control

Reading of 1.5-3.5:
- SQE Guidelines
- Basic Verification Suite
- Extensive Code Coverage Regression

Reading of 3.5-7.5 (Graded Scale:
- Most of Element Types Verified
- Most of Materials Verified
- Most of Contact Verified

Reading of 7.6-9.9 (Graded Scale):
- Most of Couplings Verified

Naturally any such Scale reading has to take into consideration the features as used for the application, regime, and even fidelity of interest. Obviously such a "Meter" is still subjective, still qualitative in how we set the meter. We should of course strive for more than a "1-10" scale. More desirable are quantitative Verification Statements such as the one given above in the DEFINITIONS.

Like the Verification or VER Meter, the concept of a Validation or VAL "Meter", with a Scale that reads

0-10, is simple but it should be fairly clear in intent. We suggest that the following factors help in setting this somewhat subjective but informative "0-10" Validation Scale:

Reading of 0-1.5:
- Runs the 1st time step
- Runs desired model to completion
- Obtains an answer: "Blind Trust" fidelity

Reading of 1.5-3.5:
- Calibrated model

Reading of 3.5-5.5:
- Mesh Convergence
- Temporal Convergence
- Code-to-Code Validation
- Sensitivities Qualitatively Correct

Reading of 5.5-7.5 (Graded Scale):
- Validation to more than one integral system
- Validation to numerous tests
- Quantitative Validation Statement
- Predictive Validation Bound Assessed

Reading of 7.5-9.9 (Graded Scale):
- All Uncertainty terms quantified
- "Fully Validated" [oxymoron?]

Naturally any such Scale reading has to take into consideration the features as used for the application, regime, and even fidelity of interest. The VAL Meter reading is as subjective as the VER Meter. The V&V process is much deeper and more quantitative than any single summary number can depict. But, the meters go beyond a "yes/no" V&V statement for communicating fidelity quickly.

Linkage between Uncertainty and Validation Level have been proposed by others as well, such the "0-1" Validation Scale of Oberkampf and Trucano[7]. When the simple quantity V=2/U is overlaid on their Validation Metric, the curves are similar. Both our "0-10" scale and other measures like the "0-1" scale of Oberkampf and Trucano[7], and in addition the "0-5" Scale of Trucano et al[8] are useful and more valuable than either no indicator at all or an oxymoron like "fully validated" or "validated code". Both expressions for "V" should be used with care.

To begin to take us beyond the simplistic "meter" or "Scale" summary, we have tried to capture four of the key levels of Validation contained in the following nomenclature:

DEMO: Demonstration, i.e. run to completion
CAL: Calibration, i.e. satisfactory agreement with accepted legacy metrics
VAL: Validation, agreement to predefined metrics, without [further] calibration
PVAL: Predictive Validation, i.e. agreement of a pre-test prediction with the test result within the pretest confidence bounds of uncertainty established through Validation.

Uncertainty Quantification (UQ):

Since we endorse the concept that validation should be a quantitative statement, it is essential to consider validation and Uncertainty Quantification (UQ) as a set. Uncertainty is inseparable from the confidence at which it is stated. Therefore, validation, uncertainty, and confidence become inseparable.

We now introduce definitions of the components of uncertainty; in this work we will focus mainly on some simple ways to handle Epistemic Uncertainty and Model Uncertainty. Consider, after Oberkampf and Trucano[3], these four components of "Error" or "Uncertainty", as defined here as the differences between quantities "q" of interest:

For a measurement of the quantity 'q' of interest, let[3]:

$$\Delta = (q_{nature}-q_{exp})+(q_{exp}-q_{exact})+(q_{exact}-q_{discrete}) \qquad [2a]$$

or alternately[3]:

$$\Delta = E_1+ E_2+ E_3+ E_4 \qquad [2b]$$

We might consider adding to Oberkampf and Trucano's set an E_0, and characterizing the E_i loosely with words, so that now:

Variability / Aleatory Uncertainty
$\mathbf{E_0=(q_{nature}-q_{nature})}$: [3a]

Uncertainty / Epistemic Uncertainty
$\mathbf{E_1=(q_{nature}-q_{exp})}$: [3b]

Error, in model
$\mathbf{E_2=(q_{exp}-q_{exact})}$: [3c]

Error, due to formulation or weak form
$\mathbf{E_3=(q_{exact}-q_{h,t>0})}$: [3d]

Error, due to discretization or solution error
$\mathbf{E_4=(q_{h,t>0}- q_{h,t,I,C})}$: [3e]

The above E_i are a useful "notional linear combination" of error / uncertainty contributions in a validated analysis. Our quantitative method will require that we express variability, uncertainty, and error as the generalized U_{Ci}, which becomes an input to the Quantitative Reliability at Confidence method. To generate each uncertainty term U_{Ci} on an environmental condition (subscript C) due to a model contribution (subscript 'i'), we combine *independent* uncertainty contributors (subscript 'j' on the Parameter uncertainty or change U_{Pj} or Δ_{Pj}) and the sensitivity S_{Cij} to them as:

$$U_{Ci}=\mathrm{RSS}(S_{Cij}U_{Pj}) \qquad [3f]$$

Here we use "RSS" to represent the root-sum-squared quadrature that we can assess for *independent* contributors. The sources, numerical values, and nature [i.e. assumed or known form of PDF, Probability Distribution Function] for all these "notional linear combination" E_i and the more rigorously combined Eqn. 3f should be stated as part of validation with uncertainty quantification. That is, enough information should be provided as part of the Validation / UQ to enable the QRC methods of this work (or similar methods) to be accomplished quantitatively and hence enable a *quantitative* validation statement.

10 steps to a V&V process: Summary Level Considerations

Our 10-step V&V summary is as follows:

1. A Program Plan should exist with scope and timeline to balance the ability to build, verify, and validate code and model capability against the assessment needs of the product line.

2. A Code Capability Plan should exist, with a simple method to track the V&V status of each capability.

3. A Risk Assessment methodology should exist and be documented. For example, in the area of engineering mechanics, a major part of our risk assessment mitigation is the ability to use multiple codes for any given analysis or planned analysis.

4. Verification (code feature) and Validation (prioritized system requirements) listings should exist and be prioritized. For Validation, an example of such a target list would be a Matrix of R_i and C_i terms for all "i" environments in our Requirements. The Validation list is of necessity tied to the prioritized product line assessment and certification needs and timing.

5. The long term plan from code features through V&V to assessment capability should address the sequence of SQE (Software Quality Engineering), VER (Verification), CAL (Calibration), and VAL (Validation):

   a. SQE: Software Quality Engineering: The software quality engineering practices may be tailored for each individual code, but should conform to a standard accepted by the developers' and users' organization. The ones we use include the recent TriLab ASCI/V&V/SQE of Hodges et al[9], and the LLNL ASCI/V&V/SQE supplement of Storch et al[10].

   b. VER_{code}: Code Verification: In the maximum state of temporal, spatial, and iterative convergence achievable, we assess the remaining error in the answers provided by the code for the feature being verified.

   c. VER_{soln}: Solution Verification: We assess the components of model error change as a function of discretization refinement in the temporal, spatial, and iterative domains. Model speed (e.g. element-steps per millisecond or inverse grind time) should be reported as it is key to determining tradeoffs of platform usage, time to solution, and quality of solution.

   d. CAL: Calibration of a Model .We establish and show existence of a model fit to some assemblage of data, but not uniqueness.

   e. VAL_{cvc}: (Risk Mitigation) Validation, Code Vs. Code (CVC). This method is often time and cost effective but fraught with dangers of misinterpretation and misuse.

   f. VAL_{suv}: (Sensitivity and Uncertainty) Validation, with Sensitivity plus Uncertainty plus Variability (SUV). This may be appropriate when for example only one integral test is available. The uncertainty bounds obtained may be quite wide, especially when several terms are rolled up (multiplied) in succession. In this VAL_{suv} method we multiply the sensitivities S_{Cij} by what may well be large and estimated material, environment, tolerancing, or other parameter uncertainties U_{Pj}. Adequate statistical quantities of test or measurement data on the U_{Pj} will help tighten our uncertainty bounds to meaningful levels on rolling up several terms. This information about each U_{Pj} must be obtained at a cost justified by its reduction in a given U_{Pj} and the importance of that U_{Pj} in total system performance. This importance is determined by knowing the values of the sensitivities S_{Cij}. We must therefore have some verification and experimental component or modular validation [see the step below and

AIAA[2], on Phased V&V] to V&V the model sensitivities S_{Cij}. This helps mitigate the interpolation and extrapolation dangers inside or outside the validated region. These derivatives S_{Cij} form the basis of our integral i^th environment uncertainty estimates for this method. For example, for the i^th environment, our Uncertainty in Capability of the system may be:

$$U_{Ci}=RSS(S_{Cij}U_{Pj}) \qquad [3f]$$

As with Verification, reporting model speed (e.g. element-steps per millisecond or inverse grind time, total model run time, and total user time) is key to determining tradeoffs of platform usage, time to solution, and quality of solution.

g. <u>VAL_{mlv}</u>: Validation at the integral level, across the rolled up set of environments of application, can be done using a Maximum Likelihood Validation (mlv) fit to the integral data. In other words, if we are fortunate to have several integral tests, the demand that our model match over all of them will more readily quantify tighter uncertainty bounds. In other words, it helps us solve the dilemma that we usually do not have enough data (or model validation) to do VAL_{suv} without the subsequent rollup of terms leading to huge uncertainty bounds. And yet, with an adequate number of integral tests, we have a body of evidence that says that the uncertainties are not so boundless as our (incomplete) VAL_{suv} would contend. VAL_{mlv} is a way of showing our likely uncertainty embedded in the integral model vs. the integral data. The Validation, if done cross-test and cross-system (several systems with a similar mission) will be more robust and with tighter quantified uncertainty bounds.

h. <u>VAL_{pre}</u>: Validation using Prediction (PRE). In principle we can "predict" tests that have already been done if we use a calibrated, validated model, with no further degrees of freedom adjusted, and then "predict" one additional pre-existing test N=N+1. However, this will always leave some doubt as to a subconscious bias of our fitting process, since we may have been influenced by that existing data even though it was not directly used to develop our validated model fit. A measure of VAL_{pre} can be obtained by using a low CAL/VAL ratio in the VAL_{mlv} process, so that if we do no additional tuning after our VAL_{mlv} fit to part of the data, we can "predict" (really post-predict) the rest of our existing data and see how good our "predictive fit" would be for the data we have.

6. Phased or Tiered V&V: After AIAA[2] and Trucano et al[8], we suggest that the previous step be denoted and tracked at one of four phases or tiers:

a. Unit Verification or Validation: A single code feature or quantity of interest

b. Benchmark or simply coupled V&V: A single coupling of features or quantities of interest; likely still a single type of physics e.g. mechanics *or* thermal *or* fluids. Verification should still be possible for this class. Validation at the component level.

c. Complex coupling V&V: Validation at the component or integral level, with multi-physics couplings e.g. mechanics *with* thermal *with* fluids. Verification will be difficult if not impossible; some Method of Manufactured Solutions (MMS) verification as in Roache[11] may be possible. (Analogous to VAL_{suv} of Step 5f).

d. Integral V&V: A V&V level suitable for integral system assessment, qualification, and certification. (Analogous to VAL_{mlv} of Step 5g).

7. <u>Metrics: Quantitative V&V, Independent of the Code Development Teams</u>:

a. Specification of a set of metrics for assessing the execution of the above activities. We suggest that when a code capability is acquired or declared by code development, a subsequent 1-2 year period is needed to enable the quantified V&V.

b. Assessment of code and model performance as determined by the ultimate users of the code product, specifically V&V and/or design and assessment teams not part of the development team, for the selected activities via the defined metrics. If adequacy or accreditation is not addressed, V&V should at least provide the quantitative information needed to address adequacy.

8. <u>Path Forward:</u> Generation of future V&V [and code development] actions based on the outcomes of this assessment; this should address any additional items needed from code development, V&V, or experimental data.

9. Documentation and archiving of experimental data and model results sufficient for future traceability and reproducibility of V&V activities.

10. Integration into the larger V&V Community.

RELATION OF V&V TO ADEQUACY, ACCEPTANCE, QUALIFICATION, AND CERTIFICATION

In order to set true acceptability Metrics for Validation, we must know where we are going – what the model tells us, and how much it matters.

We show a path from V&V (with UQ i.e. Uncertainties at stated Confidence) through the use of validated models to get component and integral system margins (i.e. Factor of Safety = Margin+1). With these quantities, we proceed to an **assessment** of Quantitative Reliability at Confidence equivalent for conditions of little or no full system testing.

An important theme is that these are the areas from which we derive "acceptance" or "adequacy" criteria. What is "good enough" V&V is a function not only of Requirements but of the implications regarding Benefit / Cost tradeoffs with time and the rest of the mission, business plan, and product line. We feel that to be ready to address "acceptance" in V&V, it is essential to have a process that leads, in the end, to a Benefit / Cost closure. In other words, our Quantitative V&V Statements have to balance both sufficiency and efficiency[6]:

Sufficiency: Doing enough V&V

Efficiency: Optimal balancing of our resources

We must prioritize our resources and determine the acceptable level of uncertainty. For product acceptance [in our case stockpile stewardship], it is our responsibility to assure, with high confidence "C" in a high assessed reliability "R", that the system can and will accommodate its lifetime of environments. We are developing and quantifying such a methodology to lead us to values of C and assessed R to fold into a total performance number or "non-frequentist RC-Equivalent" for the component or system level. We emphasize the wording, *assessed* Reliability at Confidence, because although we may never disprove an assertion that Reliability is in fact Unity, we can only quantify what we can assess with positive numerical evidence.

This Quantitative Reliability at Confidence assessment is the next step toward closure of our Value Engineering / Investment Strategy method. QRC depends on the notion that Uncertainty (U) and Confidence [C] are statistically linked; whether we are coin flipping with a frequentist number of coins "N" or using an inferred number of information points "fuzzy N*". With an (albeit assumed) PDF, we can use our V&V-obtained Uncertainty at Confidence *statistical validation statement* to obtain an assessed Reliability at Confidence. At this juncture, we note for quantified V&V:

- V&V has a statistical nature, whether with a frequentist number of comparisons N or inferential, relevance based N*
- V&V must provide uncertainty at a stated [and quantified] statistical confidence
- V&V must show the origins of its N or N* and the [perhaps expert judgment] weightings used
- V&V must allow us to assess a Reliability measure [R] from Margin (M) and assumed or known PDF, whether normal distribution or other
- V&V must *provide the information* to address adequacy, before stating whether a given model is "validated for its application" or not

These factors will allow the V&V process (and resulting validated models) to contribute to the Value Engineering Investment Strategy that has a mathematical closure (in dollars for example).

Of late, there has been much interest in quantitative certification, a quest for "confidence" that is more than just "low", "medium", or "high". It is essential to clarify what methods can and cannot be credibly used under given circumstances, because of the importance of the topic and the methods, and above all, the emerging desire to use them as **business decision and investment strategy tools.** There are several methods developed at our partner national laboratories as described and referenced more thoroughly by Logan and Nitta[5]. These works have helped to motivate our own methodology.

From V&V to QRC: Quantified Reliability at Confidence

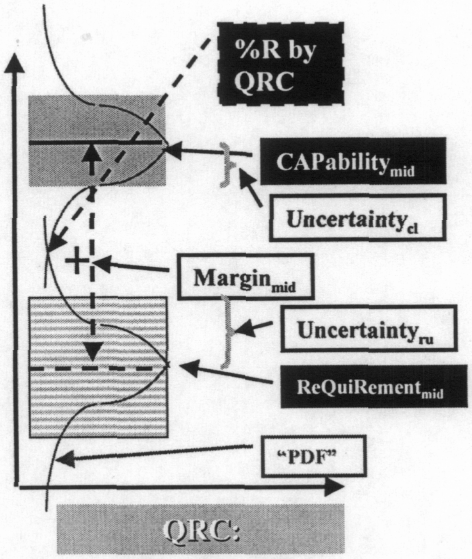
We will describe a general process for moving from V&V to quantified certification as we view it[5]. We present the basic concept for the relation between V&V and QRC or quantified reliability at confidence. We have built as much rigor and [quantified] judgment as we can into QRC, and employed statistical terms, be they frequentist or inferential in nature. We have balanced this against our desire to keep our overall equation [1] modular in nature; meaning that we can successively 'turn off' or 'set to unity' features and terms until the equations and methods reduce to the binomial coin-flipping situation. The path of V&V with UQ, then QRC, and then QSV is one way of expressing a methodology from V&V to Value Engineering with traceable, quantified closure.

Moving toward the middle of Figure 1, we can now express a quantity "M/U" which has been called Figure of Merit (FOM) in a notional sense. In this equation, "U" is a global "U" for a given environment scenario. FOM=M/U is a simple concept and indeed an old one, but it cannot be left open ended; for closure, "U" requires specification of confidence level "C". Of note is the subscript on the *second* (M/U) term (at the right of center).

We must go beyond an open-ended "M/U" in order to proceed into the realm of Quantified Reliability at Confidence (QRC). In the second (M/U) box of Fig. 1, the "Z_{qrc}=M/U at σ" after the Standard statistical Z[12], quantified uniquely and used in QRC, denotes some measure of standard deviation in a Gaussian or other Probability Distribution Function (PDF). This is an important factor required to later derive the reliability equivalent [R] quantity for the "RC Rollup" at the Tiered or Integral level. Note the presence of N*, the number of "data elements" in the set being evaluated. The quantity N* has an analogy to, for example, a binomial "N" in binomial analyses for R and C. However, N* is a weighted "N" – weighted with the relevance of number of tests, relevance of tests, number and relevance (e.g. VER and VAL settings) of analyses, or expert opinion weighting as discussed by Booker et al[4].

Measures of "M" and "U" used in QRC

Our comfort at the system level comes from the sense that "Margins" are "large". Consider Figure 3. Here, we can express our uncertainty in terms of Normal or other Probability Distribution Function (PDF) defining in essence the point where a Standard Normal Distribution variable[12] "Z" becomes Z=1. With our definition of mid-point margin "M" as M=1-FOS (Factor of Safety), there is a direct analogy between the standard statistical "Z" and our Z for QRC, Z_{qrc}=(M/U$|_{N*,\square}$). Expanding U [and de facto raising R and C as M→0] are expressed another way by Wood[13] as: "you have to express margin in terms of the probability it will be used up". The tails overlapping this closure then give us measures of R; and the method we use to obtain this information (equivalent sample size N*, our fuzzy N) gives us the value of C. For these methods, we go beyond a simple "go or no-go" Figure-Of-Merit criterion, to a statement that $|Z|>n_\sigma\ \sigma$, where n_σ is the number of standard deviations over which we define our uncertainty (for the normal PDF assumption). This way, we obtain *lower bound assessed* numerical equivalents for Reliability R at Confidence C that can be quantitatively defended from our V&V assessment process. If the resulting R at C are not acceptable, we may be able to improve them by further investment in the codes, V&V, or validation testing. These are **then investment strategy choices** we can trade off in a quantitative way as described by Logan and Nitta[5].



Z_{qrc}=M_{mid}/U$|_{N*,\sigma}$
Specify U@ eg 2σ %C
Close "PDF" tails to get %R @ %C

Figure 3. QRC's Z_{qrc} method for Margins and Uncertainties allows quantitative extension to R(M,U) at C(N*), and then to QSV. U$|_{N*,s}$ e.g. root-sum-squared quadrature.

Following are some key points about QRC, our preferred method for linking V&V and Uncertainty to Reliability and Confidence and ultimately value and investment measures:

- QRC links Margin "M" and Uncertainty "U", directly to "Reliability R at Confidence C".

- The "$Z_{qrc}=M/U|_{\sigma,N}*$" term in QRC with a mid-point M is a good start as long as physical feel for the quantities we are used to is not lost; "M/U" has been used for decades as Signal-To-Noise and Process Control. Since U at C is a direct output of our preferred validation statement, the "$Z_{qrc}=M/U|_{\sigma,N}*$" term in QRC, as opposed to some other methods, enables closure; it is not open-ended.

- Most of the components of QRC are not really new. QRC certainly builds, in our business, on the works referenced by Logan and Nitta[5] and others, just as it builds on the standard statistical definition of "Z". We believe what is new about QRC is that it enables complete and quantified (in reliability, dollars, or both) linkage from V&V with UQ, through QRC, into a Value Engineered Investment Strategy.

- QRC does in fact employ the definition of U as $U|_{\sigma,N}*$, or U at Confidence C (i.e. our quantitative validation statement).

- QRC does define Confidence C quantitatively, but with a simple analogy of our Inferred, Fuzzy N* to the frequentist N of coin-flipping.

Example of Process to a Quantitative Validation Statement

The preceding text and figures have laid the framework to take us through a Quantitative Reliability at Confidence (QRC) analysis. We now provide an example of the use of available data and model analysis results to generate the numbers necessary to make a Quantitative Validation Statement.

An example, using a flow-based model of internal combustion engine power output as a function of exhaust restriction will show how we extend the methodology from a confidence bounded uncertainty to obtain the Z_{qrc} term we need for Quantitative Reliability at Confidence, QRC. Our desire is to take some "available data" regarding power output versus exhaust restriction (translate, noise reduction), and then find the exhaust restriction "Margin"; that is, the amount of exhaust flow restriction we can tolerate and still produce the desired power output, say "300" or the value on the plot's x-axis.

But, our "available data" just looks like a pattern full of scatter (the mixed condition data points in Figure 4). This is because, as is most complex systems, power output is a function of many things – in this case not just exhaust restriction, but many other factors that were varied from test-to-test. This is not ideal in trying to isolate the effect of exhaust restriction, but it may be all the data we have. We are lucky to have as many as N=29 data points; this relatively large "N" will help reduce our epistemic uncertainty error E_1. Our goal is to use a suitable model to tell us what the power output would have been, under standard conditions with a "standard production build", with exhaust restriction then being the only variable remaining. Suffice it to say that there is such a model; in this case a "closed form" model, hence free of weak form, approximation, and spatial-temporal discretization errors, so we could claim $E_3=E_4=0$, but with many potential "model errors" or "physics errors" that we would lump into the E_2 term. In addition, this particular model provides a smooth fit (the smooth central line in Figure 4) on normalizing all N=29 data points, but it uses K=6 degrees of freedom in the model fit. These are mostly "physically based" degrees of freedom; that is, with enough component level test data we could get down to say K=1; nonetheless, since all we have is system level power output, we can only guess what values these component-level model input numbers would have been or might have been, and use them as degrees of freedom to fit out system level model of power output.

With N=29 and model degrees of freedom K=6, our closed-form model does quite well in providing a "smooth curve" of output versus restriction. Some method for UQ, Uncertainty Quantification, must be used to tell us how well our model really did capture the data as measured, else we would have no "confidence" in using this model to provide a curve for "standard condition" performance. The outer (2σ) Confidence Bounded Uncertainties in Figure 4 account for the difference between the actual "mixed condition" data available and our model fit to that data, we can establish model error (including bias error as the reader may notice) in our model. As we might expect, the confidence bound estimates are broad, in part due to our epistemic uncertainty E_1, and they become more broad as we get more distant from where the actual measurements are clustered, and we can observe a bias error in the model result as the confidence bounds appear skewed.

We note, as shown in Figure 4, that even though there was a lot (N=29 points) of "available data", it will not line up with our curve for standard condition output vs. restriction. It just means that our model was built from a lot of data (N=29 points). Our confidence in the model is expressed by the bounding lines. The fact that the raw available data does not fit our smooth "standard condition" line is in fact an indication that we are extrapolating something – not restriction, but some other parameter affecting output at a given restriction – to get to standard conditions. Or, it may be that one of our raw data points happens to lie right on the standard condition model line; this may mean we did not extrapolate at all, or it may mean that we used our model to extrapolate two non-standard conditions and ended up "back on the curve, so it *looks like* we did not extrapolate at all.

This methodology and numerics are a minimum set for a minimal fidelity QRC analysis. Next, we sketch some quantities of interest on our model for output to capture our QRC measure of Margin and Uncertainty, "$Z_{qrc}=M/U|_{N^*,\sigma}$". If we now compare Figure 4 to Figure 3, we see the similarity in the quantities "mid-point M" (that is, power capability P_c minus power required P_r), and the relevant model uncertainties (E_1 and E_2), lower on capability U_{cl}, and upper on requirement U_{ru}; we assume $U_{ru}=0$ in this example (See Figure 3). Now, if there is no further uncertainty, ie aleatoric uncertainty error $E_l=0$ and approximation / discretization / solution uncertainty errors $E_3=E_4=0$ (a good approximation in this closed-form case). Then our assessed lower bound reliability "R" is the number of Confidence-corrected sigmas defined by this specific "$Z_{qrc}=M/U|_{N^*,\sigma}$".

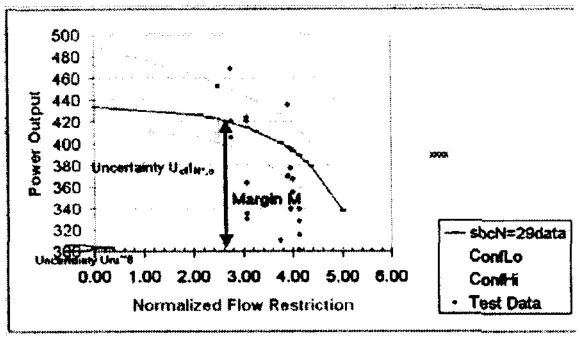


Figure 4. Confidence-bounded (e.g. 2-sigma) standard-condition model fit to available mixed-condition data. Final steps include assessment of Margin, enabling determination of Z_{qrc}.

Now that we have worked our example all the way through Figure 4, we can proceed to use the QRC value we have generated. However, in this case, our VAL_{mlv} (Maximum Likelihood) fit only makes inferences about individual sensitivities and parameter input uncertainties based on our integral fit to the system level data. We cannot directly tell which parameter or sensitivity contributed the most to our confidence-bounded uncertainty band. A *model based* sensitivity table should be generated before significant funds are spent on validation component or system tests.

The QRC value enters into the Risk Diagram of Fig. 2. As the level of V&V goes up, our Likelihood of Risk (due to poor model V&V) goes down. Multiplying this V&V-QRC based likelihood by the separately determined Consequence in Fig. 2 provides a direct (dollar benefit) measure of V&V. This is shown in Fig. 2, where *adequate* V&V reduces the *likelihood* of failure, and *timely* V&V (i.e. concurrent with design vs. post-mortem) can further reduce Risk if we design our business and products to minimize the *consequence* as well.

SUMMARY and CONCLUSIONS:

Our V&V methodology includes a process leading to quantitative V&V statements; that is, confidence bounded uncertainty on a quantity of interest over a specified regime of interest. The concept of a simple "0-10 Scale" is suggested as a *first step* toward quantitative V&V. Quantitative statistically based V&V statements will be the next evolutionary step.

The advantage of Z_{qrc}, leading to Quantified Reliability "R" at Confidence "C" (QRC) is that it lets us:

- Relate Margin "M" and Uncertainty "U" to "R".
- Demand that we associate "R" with a "C" - and quantifies that "C" based on eg "Fuzzy N*".
- Show how better assessments of M and U - and increasing the "effective number of coin flips" N* - quantitatively tightens U allowing higher quantified C.
- Eventually, as shown by Logan and Nitta[5], we can express this situation as value *(dollars)* via QSV - and so we can ***protect*** **the investments in computing, assessments, tests,** etc by *Quantifying* their *System Value (QSV)*. This method can provide a clear link between "science" and V&V and "Value" and "Investment Strategy" – *it is our hope that QRC-into-QSV will provide this link.*
- Provide a numerical [albeit judgment folded] estimate of "how much confidence do we need and how do we get it".

- Provide a quantitative V&V statement as a lower statistical bound. This has the advantage of being *quantitative and assessable*, but also the advantage in that we recognize that the upper bound on a product may still be quite high (even unity); this avoids conflict with assertions, whatever the source, that “high reliability is expected or promised” from a given product. The V&V-based lower bound *assessment* and the upper bound *assertion* can both be right; and both have appropriate uses.

ACKNOWLEDGEMENTS:

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract W-7405-Eng-48. The authors have had many helpful discussions with many colleagues in this rapidly evolving area. We would especially like to acknowledge helpful discussions with our partner lab colleagues. Discussions and insight from Bill Oberkampf, Tim Trucano, and Marty Pilch of Sandia have motivated our work. We also had many helpful discussions and concepts contributed by Cheryl Faust, Jane Booker, Mike Hamada, Ron Dolin, David Sharp, Karen Pao, Merri Wood, Peter Adams, Dawn Flicker and Scott Carman, of Los Alamos and Livermore respectively. These colleagues at our partner labs are pushing the barriers of the quantitative linkages between validation, uncertainty, and confidence in the stockpile and it is our privilege to interact with them. We also would like to thank Carolyn Mangeng, Bill Reed, and Rich Wagner for encouragement to pursue these concepts and this work.

NOMENCLATURE:

ASCI	Accelerated Strategic Computing Initiative
B	Benefit, usually in $$$, Millions ($M), or Billions ($B)
BCR	Benefit / Cost Ratio
C (In BCR)	Cost, usually in $$$, Millions ($M), or Billions ($B)
C_i (In QRC)	Confidence, in the ith environment (i omitted if only one)
E_i	Error or Uncertainty Components
FOM	Figure of Merit, "M/U"
LLNL	Lawrence Livermore National Laboratory
M	Margin, where Factor of Safety = M+1
M_i	Margin, in the ith environment
μ	population mean
N	Frequentist N, Number of trials as in coin-flipping
N*	Inferred, Weighted, Bayesian or Fuzzy N
PDF	Probability Distribution Function
QRC	Quantitative Reliability at Confidence
QSV	Quantitative Stockpile Value
PVF	Present Value Factor, 0<PVF<1
R_i	Reliability in the ith environment (i omitted if only one)
RC*	Reliability at Confidence product equivalent
S_{Cij}	Sensitivity in Capability in the ith environment to the jth parameter
σ	population standard deviation
U	Uncertainty, General or "System" [in V&V always at a confidence C]
U_{Ci}	Uncertainty in Capability for the ith environment
U_{Pi}	Uncertainty in Parameter for the jth parameter [material, tolerance, etc.]
UQ	Uncertainty Quantification
V&V	Verification & Validation
Z	Standard Normal Distribution Variable for variable X, $Z=(X-\mu)/s$
Z_{qrc}	Standard Normal Distribution Variable, $Z=M/U|_{N^*,\sigma}$ in QRC

REFERENCES:

[1] J.A. Cafeo and P.J. Roache, private communication of draft V&V definitions, April, 2002.

[2] AIAA, “Guide for the Verification and Validation of Computational Fluid Dynamics Simulations”, AIAA-G-077-1998, Reston, VA, 1998.

[3] W.L. Oberkampf and T.G. Trucano, “Verification and Validation in Computational Fluid Dynamics”, SAND2002-0529, March 2002.

[4] J.M. Booker, T.R. Bement, and K.F. Sellers, "Linking Probability Theory and Fuzzy Sets – A Study in Uncertainty Assessment", AIAA-2002-1570, April, 2002.

[5] R.W. Logan and C.K. Nitta, "Verification & Validation (V&V) Methodology and Quantitative Reliability at Confidence (QRC): Basis for an Investment Strategy", LLNL UCRL-ID-150874, November, 2002.

[6] M. Pilch, private communication, 28 October, 2002.

[7] W.L. Oberkampf and T.G. Trucano, "Validation Methodology in Computational Fluid Dynamics", AIAA 2000-2549, Invited Paper at Fluids 2000, Denver, CO, June 2000.

[8] T.G. Trucano, M. Pilch, and W.L. Oberkampf, "General Concepts for Experimental Validation of ASCI Code Applications", SAND-2002-0341, March 2002.

[9] A. Hodges, G. Froehlich, D. Peercy, M. Pilch, J. Meza, M. Peterson, J. LaGrange, L. Cox, K. Koch, N. Storch, C. Nitta, and E. Dube, "ASCI Software Quality Engineering, Goals, Principles, and Guidelines", DOE/DP/ASC-SQE-2000-FDRFR-VERS2, February, 2001.

[10] N. Storch, C. Nitta, R. Klein, T. Quinn, S. Louis, D. Sam, E. Dube, L. Cook, D. Miller, and P. Miller, "LLNL Site-Specific ASCI Software Quality Engineering Recommended Practices", Version 1.0, LLNL UCRL-ID-143698, May, 2001.

[11] P.J. Roache, *Verification and Validation in Computational Science and Engineering*, Hermosa Publishers, Albuquerque, NM.

[12] D.S. Moore and G.P. McCabe, *Introduction to the Practice of Statistics*, W.H. Freeman & Co., 1989.

[13] M.M. Wood, private communication, 16 November, 2000.